MANIFOLD BLURRING MEAN SHIFT ALGORITHMS FOR MANIFOLD DENOISING.

Weiran Wang and Miguel Á. Carreira-Perpiñán.
EECS, School of Engineering, University of California, Merced.

1 Abstract

We propose a new family of algorithms for denoising data assumed to lie on a low-dimensional manifold. The algorithms are based on the blurring mean-shift update, which moves each data point towards its neighbors, but constrain the motion to be orthogonal to the manifold. The resulting algorithms are nonparametric, simple to implement and very effective at removing noise while preserving the curvature of the manifold and limiting shrinkage. They deal well with extreme outliers and with variations of density along the manifold. We apply them as preprocessing for dimensionality reduction; and for nearest-neighbor classification of MNIST digits, with consistent improvements up to 36% over the original data.

4 How to set the parameters?

- k: the number of nearest neighbors that estimates the local tangent space. It typically grows sublinearly with N.
- L: local intrinsic dimension. It could be estimated (e.g. using the correlation dimension) but here we fix it. Note $L = 0$ is GBMS, $L = D$ is no motion.
- σ: related to the level of local noise outside the manifold. The larger σ, the stronger the denoising and the distortion of the manifold shape. Using a k-nn graph limits the motion to near the k nearest neighbors and allows larger σ. Note $\sigma = 0$ is no motion, $\sigma = \infty$ is LTP.

5 Convergence results

- The MBMS algorithms are covariant under rigid motions.
- Any dataset that is contained in an L-dim. linear manifold is a fixed point of MBMS (since the tangent space coincides with this manifold and tangential motion is removed).
- A Gaussian distribution converges cubically to the linear manifold defined by its mean and L leading eigenvectors (denoising proceeds independently along each principal axis but motion along the L leading eigenvectors is removed). Essentially, MBMS performs GBMS clustering orthogonal to the principal manifold.

6 Stopping criteria

A practical indicator of whether we have achieved significant denoising while preventing shrinkage is the histogram over all data points of the orthogonal variance $\lambda_\perp$ (the sum of the trailing $k - L$ eigenvalues of $\mathbf{x}_n$'s local covariance). Its mean decreases drastically in the first few iterations, while the mean of the histogram of the tangential variance $\lambda_\parallel$ stabilizes.

7 Computational complexity

$\mathcal{O}(N^2D + N(D+k)\min(D,k)^2)$ per iteration, where the first term is for finding nearest neighbors and for the mean-shift step, and the second for the local PCAs. Various accelerations possible, e.g. not updating the k-nn graph affects the result little and makes the cost linear on N.

9 Conclusion

With adequate parameter values, the proposed MBMS algorithm is very effective at denoising in a handful of iterations a dataset with low-dim. structure, even with extreme outliers, and causing very small manifold distortion. It is nonparametric and deterministic (no local optima); its only user parameters (L, k, σ) are intuitive and good regions for them seem easy to find. LTP (local tangent projection) is a particular, simple case of MBMS that has quasi-optimal performance and only needs L and k. Preprocessing with MBMS improves the quality of algorithms for manifold learning and classification that are sensitive to noise or outliers. We expect MBMS to help in other settings with noisy data of intrinsic low dimensionality, such as density estimation, regression or semi-supervised learning.

Work supported by NSF CAREER award IIS–0754089.

2 The MBMS algorithm

- Local clustering with Gaussian blurring mean shift (GBMS): the blurring mean-shift update moves datapoints to the kernel average of their neighbors:

$$\mathbf{x}_n \leftarrow \sum_{m \in \mathcal{N}_n} \frac{G_\sigma(\mathbf{x}_n, \mathbf{x}_m)}{\sum_{m' \in \mathcal{N}_n} G_\sigma(\mathbf{x}_n, \mathbf{x}_{m'})} \mathbf{x}_m.$$

The average is over $\mathcal{N}_n = \{1, \dots, N\}$ (full graph) or the k nearest neighbors of $\mathbf{x}_n$ (k-nn graph), and $G_\sigma(\mathbf{x}_n, \mathbf{x}_m) \propto \exp\left(-\frac{1}{2}(\|\mathbf{x}_n - \mathbf{x}_m\|/\sigma)^2\right)$. This step produces denoising.

- Local tangent space estimation with PCA: local PCA gives the best linear L-dim. manifold in terms of reconstruction error (i.e., orthogonal projection on the manifold):

$$\min_{\boldsymbol{\mu}, \mathbf{U}} \sum_{m \in \mathcal{N}'_n} \left\| \mathbf{x}_m - (\mathbf{U}\mathbf{U}^T(\mathbf{x}_m - \boldsymbol{\mu}) + \boldsymbol{\mu}) \right\|^2$$

s.t. $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ with $U_{D\times L}$, $\boldsymbol{\mu}_{D\times 1}$, whose solution is $\boldsymbol{\mu} = \mathrm{E}_{\mathcal{N}'_n}\{\mathbf{x}\}$ and $\mathbf{U}$ = the leading L eigenvectors of $\mathrm{cov}_{\mathcal{N}'_n}\{\mathbf{x}\}$. In general, $\mathcal{N}_n$ need not equal $\mathcal{N}'_n$. This step maintains the manifold structure.

3 Varieties of the MBMS algorithm

MBMS (L, k, σ) with full graph or k-nn graph: given $\mathbf{X}_{D\times N}$

repeat
  for $n = 1, \dots, N$
    $\mathcal{N}_n \leftarrow \{1, \dots, N\}$ (full graph: MBMSf) or
      k nearest neighbors of $\mathbf{x}_n$ (k-nn graph: MBMSk)
    $\partial\mathbf{x}_n \leftarrow -\mathbf{x}_n + \sum_{m\in\mathcal{N}_n} \frac{G_\sigma(\mathbf{x}_n,\mathbf{x}_m)}{\sum_{m'\in\mathcal{N}_n} G_\sigma(\mathbf{x}_n,\mathbf{x}_{m'})}\mathbf{x}_m$ — mean-shift step
    $\mathcal{X}_n \leftarrow k$ nearest neighbors of $\mathbf{x}_n$
    $(\boldsymbol{\mu}_n, \mathbf{U}_n) \leftarrow \mathrm{PCA}_L(\mathcal{X}_n)$ — estimate L-dim tangent space at $\mathbf{x}_n$
    $\partial\mathbf{x}_n \leftarrow (\mathbf{I} - \mathbf{U}_n\mathbf{U}_n^T)\partial\mathbf{x}_n$ — subtract parallel motion
  end
  $\mathbf{X} \leftarrow \mathbf{X} + \partial\mathbf{X}$ — move points
until stop
return $\mathbf{X}$

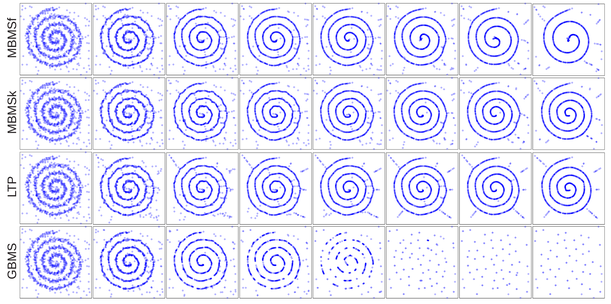
LTP (L, k) with k-nn graph: given $\mathbf{X}_{D\times N}$

repeat
  for $n = 1, \dots, N$
    $\mathcal{X}_n \leftarrow k$ nearest neighbors of $\mathbf{x}_n$
    $(\boldsymbol{\mu}_n, \mathbf{U}_n) \leftarrow \mathrm{PCA}_L(\mathcal{X}_n)$ — estimate L-dim tangent space at $\mathbf{x}_n$
    $\partial\mathbf{x}_n \leftarrow (\mathbf{I} - \mathbf{U}_n\mathbf{U}_n^T)(\boldsymbol{\mu}_n - \mathbf{x}_n)$ — project point onto tangent space
  end
  $\mathbf{X} \leftarrow \mathbf{X} + \partial\mathbf{X}$ — move points
until stop
return $\mathbf{X}$

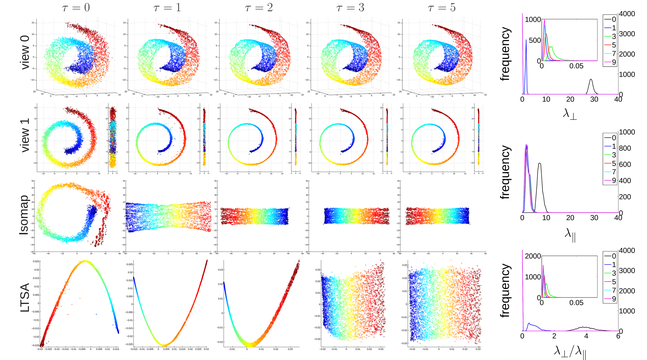
GBMS (k, σ) with full or k-nn graph: given $\mathbf{X}_{D\times N}$

repeat
  for $n = 1, \dots, N$
    $\mathcal{N}_n \leftarrow \{1, \dots, N\}$ (full graph) or
      k nearest neighbors of $\mathbf{x}_n$ (k-nn graph)
    $\partial\mathbf{x}_n \leftarrow -\mathbf{x}_n + \sum_{m\in\mathcal{N}_n} \frac{G_\sigma(\mathbf{x}_n,\mathbf{x}_m)}{\sum_{m'\in\mathcal{N}_n} G_\sigma(\mathbf{x}_n,\mathbf{x}_{m'})}\mathbf{x}_m$ — mean-shift step
  end
  $\mathbf{X} \leftarrow \mathbf{X} + \partial\mathbf{X}$ — move points
until stop
return $\mathbf{X}$

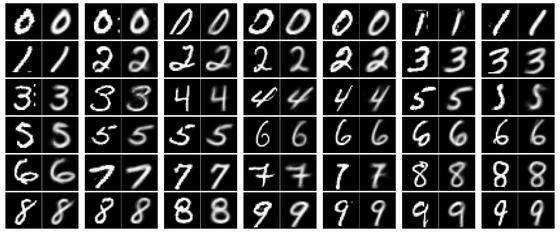
8 Experimental results



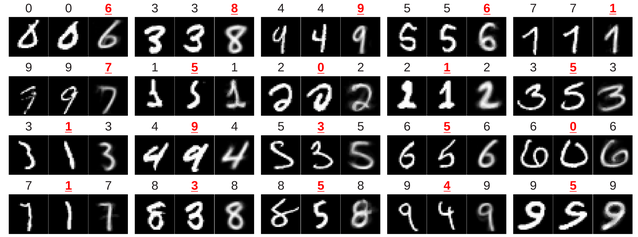
Denoising a spiral with outliers over iterations ($\tau = 0$ is the original dataset). Each box is the square $[-30, 30]^2$, where 100 outliers were uniformly added to an existing 1000-point noisy spiral. Algorithms (L, k, σ): $(1, 10, 1.5)$ and full graph (MBMSf), $(1, 10, 1.5)$ and k-nn graph (MBMSk), $(1, 10, \infty)$ and k-nn graph (LTP), and $(0, \cdot, 1.5)$ and full graph (GBMS).



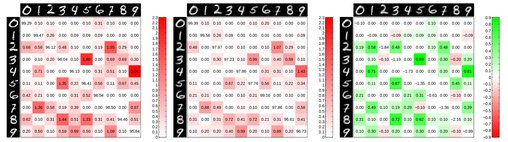
Dimensionality reduction with Isomap and LTSA for different iterations of MBMSk denoising (10–nearest-neighbor graph, $L = 2$, $k = 30$, $\sigma = 5$). $\tau = 0$ is the original Swiss roll dataset ($N = 4000$ points) lifted to 100 dimensions with additive Gaussian noise of stdev 0.6 in each dimension. Isomap/LTSA used a 10-nn graph. Isomap's residual variances ($\tau = 0, 1, 2, 3, 5$): 0.3128, 0.0030, 0.0002, 0.0002, 0.0003. Right column: histograms over all data points of the normal, tangential, and normal/tangential ratio of the variances (curves for iterations $\tau = 0, 1, 3, 5, 7, 9$).



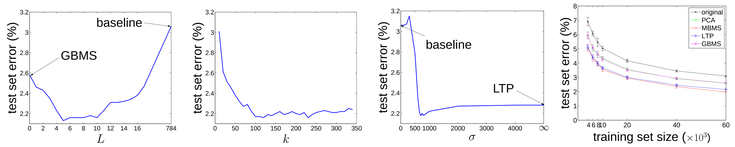
Classification of MNIST handwritten digits Sample pairs of (original,denoised) images from the training set. The digits look smoother (as if they had been anti-aliased to reduce pixelation) and easier to read, and show sophisticated "corrections" (e.g. inpainting and erasure). Comparing the original 0 1 2 3 4 5 6 7 8 9 vs the denoised 0 1 2 3 4 5 6 7 8 9, one sees this would help classification.



Some misclassified images. Each triplet is (test,original-nearest-neighbor,denoised-nearest-neighbor) and the corresponding label is above each image, with errors underlined. After denoising there are fewer errors, some of which are arguably wrong ground-truth labels.



The confusion matrices (before denoising, after denoising, before minus after). Each row shows the probability (in percentage) of recognizing each digit to all digits. On the left two matrices, white means zero error (perfect classification) and red means positive error. On the right matrix, white means no change (same error before and after denoising), green means denoising reduced the error, and red means denoising increased the error.



Left 3 plots: 5–fold cross-validation error (%) curves with a nearest-neighbor classifier on the entire MNIST training dataset (60k points, thus each fold trains on 48k and tests on 12k) using MBMSk; we selected $L = 9$, $k = 140$, $\sigma = 695$ as final values. *Right plot*: denoising and classification of the MNIST test set (10k points), by training on the entire training set (rightmost value) and also on smaller subsets of it (errorbars over 10 random subsets). Algorithms (L, k, σ), all using a k-nn graph: MBMSk $(9, 140, 695)$, LTP $(9, 140, \infty)$, GBMS $(0, 140, 600)$, and PCA ($L = 41$).